DOCUMENT RESUME

ED 409 364 TM 026 902

AUTHOR Schnipke, Deborah L.; Pashley, Peter J.
TITLE Assessing Subgroup Differences in Item Response Times.
PUB DATE Mar 97
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Assisted Testing; Data Analysis; English; *Item Bias; Language Dominance; Performance Factors; Regression (Statistics); Research Methodology; *Responses; Test Items; *Timed Tests
IDENTIFIERS High Stakes Tests; Item Bias Detection; *Subgroups; *Survival Analysis

ABSTRACT

Differences in test performance on time-limited tests may be due in part to differential response-time rates between subgroups, rather than real differences in the knowledge, skills, or developed abilities of interest. With computer-administered tests, response times are available and may be used to address this issue. This study investigates procedures for identifying possible subgroup differences in response times from a computer-administered test. Based on the results, it appears that survival analysis methodology is useful for uncovering subgroup differences in response time rates. Data from a national high-stakes computer-administered (but not adaptive) reasoning test were used. Data from 6,306 test takers who said that English was their best language and 462 test takers who said that English was not their best language were analyzed with parametric (Cox regression) and nonparametric (Wilcoxon test) approaches. Significant differences between response times across the two subgroups were found, although this does not necessarily imply that particular items should have been identified as having differential item functioning. (Contains 3 tables, 3 figures, and 11 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 409 364

Assessing Subgroup Differences in Item Response Times

Deborah L. Schnipke
Peter J. Pashley

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Deborah Schnipke
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM026 902

This paper is prepared for the Annual Meeting of the American Educational Research Association in Chicago, IL

BEST COPY AVAILABLE

Assessing Subgroup Differences in Item Response Times[1]

Deborah L. Schnipke and Peter J. Pashley
Law School Admission Council

Abstract: Differences in test performance on time-limited tests may be due in part to differential response-time rates between subgroups, rather than real differences in the knowledge, skills, or developed abilities of interest. With computer-administered tests, response times are available and may be used to address this issue. The present study investigates procedures for identifying possible subgroup differences in response times from a computer-administered test. Based on the results, it appears that survival analysis methodology is useful for uncovering subgroup differences in response time rates.

The potential always exists for subgroup differences in test performance to be attributable to the content, wording, or other irrelevant aspects of particular items, rather than real differences in the knowledge, skills, or developed abilities of interest. Several methods of differential item functioning (DIF) have been developed to statistically investigate these differences. The most popular is the Mantel-Haenzsel procedure (Holland & Thayer, 1988). Other procedures include standardization (Dorans & Kulick, 1986), SIBTEST (Shealy & Stout, 1993), and logistic discriminant function analysis (Miller & Spray, 1993). All of these procedures identify items that are relatively more difficult for some subgroups than would be expected from their performance on other items in the test.

Differences in test performance may also be due, in part, to differential response-time rates between subgroups. Differences in response-time rates of items administered early in the test may not affect item-score performance on those items, but may cause the overall form to be excessively time-consuming (speeded) for some subgroups. Because of strict time limits, excessively time-consuming forms may lead to lower test scores because test takers simply do not have enough time to work on every item. Recent studies that considered not-reached items as an indication of the effect of time limits on tests (i.e., speededness) have found, for example, that African American and Hispanic test takers tend to reach slightly fewer items than Caucasian test takers (e.g., Dorans, Schmitt, & Bleistein, 1992; O'Neill & Powers, 1993; Schaeffer, Reese, Steffen, McKinley, & Mills, 1993). Unfortunately, the number of not-reached items is typically

TM027902

[1]Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL, March, 1997.

a poor measure of speededness, especially when number-correct scoring is implemented. A more accurate measure of speededness would require an investigation of item response times (e.g., Schnipke, 1995, Schnipke & Scrams, *in press*).

Response-time analyses have shown that some subgroups, for example, Hispanics (Llabre & Froman, 1987) and African Americans (O'Neill & Powers, 1993), tend to spend more time on each item than other subgroups. Additionally, some subgroups, based on test-taker ability (Schaeffer et al., 1993) or ethnicity (Llabre & Froman, 1987), are less effective than other subgroups at allocating their time per item according to the difficulty of the item. These studies suggest that time limits on tests may differentially affect test scores of some subgroups. Unfortunately, the statistics used in these studies (i. e., means, medians) are rudimentary and do not capture the complex nature of response times.

The goal of this study is to investigate methodology for assessing subgroup differences in response-time rates. Rather than using just the average item response time (mean or median) for various subgroups, we use the more powerful approach of looking at the entire distribution of response times for each item using survival analysis methodology. This methodology is applied to data from a national, high-stakes computer-administered test.

Additionally, we investigate whether certain subgroups spend more time on items at the beginning of the test, and if this is associated with those subgroups either running out of time more so than other test takers or simply spending less time on the later, presumably more difficult items. To determine if test takers ran out of time, we will look at whether they reached items at the end of the test or suddenly responded very quickly (what Schnipke, 1995, and Schnipke & Scrams, *in press*, called rapid-guessing behavior).

Method

Instrument Used

Data from a national, high-stakes computer-administered reasoning test were used in the present study. The items were not administered adaptively; all test takers received the same 25 items. Several demographic variables were collected. One demographic variable asked whether English is the person's best language. For convenience this variable will be called English fluency. Data were available for 6,306 test takers who said that English is their best language

and for 462 test takers who said that English is not their best language. Test takers who responded that English is their best language will be referred to as primary speakers, and test takers who responded that English is not their best language will be referred to as nonprimary speakers.

Analyses

Survival analysis techniques: Cox regression. Techniques of survival analysis will be used to analyze the response times. The survival distribution function (SDF), or S(t), is the probability at time t that a test taker has not yet responded to the item (their response time denoted by T) and is given by S(t)=Prob(T>t). The SDF is closely related to the cumulative distribution function (CDF) which is defined as 1-S(t). Cox regression is a parametric approach that can be used to take into account covariates (such as test score) that influence the time variable (e.g., item response time). Because primary and nonprimary speakers have different test-score distributions on the test and this may influence response times, test score will be used as a covariate. To determine whether the two subgroups (primary and nonprimary speakers) differ in response time, the variable (English fluency) is also included in the regression.

Survival analysis techniques: Gehan's generalized Wilcoxon test. Gehan's generalized Wilcoxon test (Lee, 1992) is a nonparametric test that compares two or more survival distributions. It tests whether two or more subgroups could have arisen from identical survival functions. This technique does not control for covariates, so another way of accounting for test-score differences was needed. To control for differences in test scores in the two subgroups, a sample of test takers was randomly selected from the primary speakers to match the score distribution of the nonprimary speakers, resulting in 462 test takers in each subgroup. Thus this analysis removes the effect of score by matching test takers on this variable, then determines whether ability-matched primary and nonprimary speakers have different response time distributions.

Chi-squared analysis of unreached items and rapid-guessing behavior. Previous experience with data such as these (Schnipke, 1995; Schnipke & Scrams, *in press*) indicate that toward the end of a test, some test takers may begin to respond very quickly, faster than one could reasonably fully read and consider an item. To determine if one subgroup engaged in

rapid-guessing behavior more than the other, a χ^2 test was performed. A two-by-two table was created for each item: primary or nonprimary speaker by rapid-guessing behavior or solution behavior (responses made by test takers actively trying to find the solution to the item).

A technique suggested by Schnipke and Scrams (*in press*) was used to identify responses that most likely reflect rapid guessing. Response-time distribution functions for all items were fit with two alternative models: a single-state model and a two-state model. The single-state model is a lognormal distribution function. This model should fit the unimodal, positively skewed distributions typical of response times. The two-state model is a mixture of two lognormal distribution functions.

If some test takers are engaging in rapid guessing, their response times will be unusually fast, and the resulting density function will be bimodal. The two-state model allows for test takers engaging in two different response strategies that result in different response-time distributions. This model can be used to deconvolve the observed response-time distribution into two component distributions: one for test takers actively trying to solve the item, and one for test takers engaging in rapid-guessing behavior. The model is not deterministic, so there is no way of knowing for certain which responses resulted from each strategy. Instead, responses are assigned a likelihood ratio that compares the estimated probability that the response reflects rapid guessing to the estimated probability that it reflects an active attempt to solve the item. If the likelihood ratio is greater than 1, the response is more likely to reflect rapid guessing than active solution. All responses with a likelihood ratio greater than 1 were classified as rapid guesses. All other responses were classified as solution behavior. A sample bimodal item-response-time density function is shown in Figure 1.

Results

Cox regression. Table 1 shows the results of the Cox regression analysis. Not surprisingly, test score was a significant predictor of response times for all items. English fluency was a significant predictor on about half of the items. On the first half of the test, nonprimary speakers tended to respond slower on average than primary speakers when there was a difference (on only one item, item 9, did nonprimary speakers respond faster). On the last half

of the test, nonprimary speakers responded faster on average than primary speakers on all items where there was a difference (items 17-23).

Gehan's generalized Wilcoxon test. Figure 2 shows the cumulative distribution function (CDF) for item 1 by subgroup (primary and nonprimary speakers), using the sample of test takers matched on test scores as described above. The nonprimary speakers (solid line) tended to respond more slowly on this item (their CDF is shifted to the right). The difference between the response times for the two subgroups was statistically significant at $\alpha = 0.01$ using Gehan's generalized Wilcoxon test (Lee, 1992). This is consistent with the results of the Cox regression shown in Table 1. Figure 3 shows the CDFs and results of Gehan's generalized Wilcoxon test for the rest of the items (2-25). The two methods (Cox regression with test score as a covariate and Gehan's generalized Wilcoxon test with the test-score matched test takers) produced the same result on most items. On items 9, 13, 18, 19, and 20, the Cox regression indicated that English fluency is a significant predictor of response times, whereas the generalized Wilcoxon test indicated that English fluency was not a significant predictor of response times.

Chi-squared tests of unreached items and rapid-guessing behavior. To investigate whether nonprimary speakers were affected more by the time limit on this test, we examined the proportion of test takers from each subgroup who did not reach each item (Table 2), as well as the proportion of test takers who engaged in rapid-guessing behavior (as described above) on each item (Table 3). As shown in Table 2, a higher proportion of nonprimary speakers did not reach several items, but the differences were very small, even when the differences were statistically significant. (The minimum expected frequency was less than 5 for items 1-16, so the χ^2 results are not reliable and are not shown.)

As shown in Table 3, nonprimary speakers were more likely to engage in rapid-guessing behavior on items 9 and 18-24. These items correspond fairly well with the items on which nonprimary speakers responded faster on average than primary speakers according to the Cox regression (Table 1). The higher proportion of rapid guesses by the nonprimary speakers may explain why their response-time distribution was faster on average than the primary speakers' distribution on these items.

Discussion

Based on these results, it appears that survival analysis methodology is useful for uncovering subgroup differences in response times. In this study, nonprimary speakers responded slower on some items at the beginning of the test and faster on some items toward the end of the test (where the items tended to be more difficult). This test is known to be speeded, and the differences suggest that the nonprimary speakers may be affected more by the speed component than the primary speakers. The nonprimary test takers took more time on several items early in the test, and the strict time limit (i.e., speededness) may have caused them to rush at the end of the test. This was indicated by the higher proportion of rapid guesses by nonprimary speakers.

The analyses conducted in this study were both easy to conduct and to understand. Many statistical packages, such as SAS and SPSS, provide the types of survival analyses described here. While the rapid-guessing behavior estimation routines are not as easily implemented, they are at least easily understood. The graphical approach used here should be especially helpful to test developers who are attempting to construct fair and equitable assessments.

Parametric (Cox) and non-parametric (Wilcoxon) approaches were included in the study because, as is usually the case, both have their strengths and weaknesses. We assumed a lognormal distribution for response times when conducting the Cox analyses. While this is a reasonable choice, one could also have chosen the Weibull distribution, for example, and obtained slightly different results. The Wilcoxon approach avoids this problem, but at the price of not being able to handle covariates in a convenient manner. The latter problem was addressed by analyzing matched samples in the Wilcoxon analyses. Unfortunately, this is not always possible to do and, in any case, does not control for covariates as well as does taking them into account directly in the model specification (as in the Cox approach).

Another advantage of the survival analysis approach is that censored values can be easily accommodated. While the data used in this study contained no censored values, test takers often consider items without answering them, especially when formula scoring is employed. In such cases, although we do not know exactly how long the test taker would have taken to answer an

item, we do know it would have been more then a certain time period – information that can and should be used.

For a number of the items, significant differences between response times across the two subgroups investigated were found. This does not necessarily imply that these should have been identified as DIF items. First, we would naturally expect test takers with fewer English skills to take longer answering items. Second, as with most standard DIF procedures, the effect size should be taken into consideration. Although it was not done explicitly in this study, one could examine the survival curves produced to determine whether any preset difference levels between response-time curves have been exceeded.

Finally, in an operational DIF analyses, flagged items are always reviewed by test specialists who determine whether the effects observed are truly problematic or could more reasonably be attributed to statistical error. While not done in this study, this practice should, of course, also accompany any analyses of response times. The reasons behind certain outcomes, such as why item 9 was found to have been answered more quickly on average by nonprimary than primary speakers, could possibly be answered.

References

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*, 309-319.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.) *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Lee, E. T. (1992). *Statistical methods for survival data analysis*. New York: John Wiley & Sons, Inc.

Llabre, M. M., & Froman, T. W. (1987). Allocation of time to test items: A study of ethnic differences. *Journal of Experimental Education, 55*(3), 137-140.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*, 107-122.

O'Neill, K., & Powers, D. E. (1993, April). *The performance of examinee subgroups on a computer-administered test of basic academic skills*. Paper presented at the meeting of the National Council on Measurement in Education, Atlanta.

Schaeffer, G. A., Reese, C. M., Steffen, M., McKinley, R. L., & Mills, C. N. (1993). *Field test of a computer-based GRE general test* (Research Report No. RR-93-07). Princeton, NJ: Educational Testing Service.

Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco.

Schnipke, D. L., & Scrams, D. J. (*in press*). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*.

Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Table 1. Results of Cox Regression analysis

	Regression Estimates			
Item Number	**Intercept***	**Score***	**English Fluency****	**Faster Group****
1	4.877	-0.031	0.107	Primary speakers
2	4.720	-0.023	-0.038	n.a.
3	3.280	0.035	0.125	Primary speakers
4	3.488	-0.002	0.041	n.a.
5	4.117	0.020	-0.006	n.a.
6	3.882	-0.010	-0.019	n.a.
7	4.729	-0.029	0.087	Primary speakers
8	4.484	-0.025	0.019	n.a.
9	4.578	-0.011	-0.064	Nonprimary speakers
10	4.940	0.004	0.063	Primary speakers
11	4.116	-0.017	-0.007	n.a.
12	3.681	0.041	-0.019	n.a.
13	3.612	0.031	-0.019	Nonprimary speakers
14	3.452	0.062	-0.021	n.a.
15	4.307	0.031	-0.013	n.a.
16	3.191	0.021	-0.097	n.a.
17	3.080	0.046	-0.163	Nonprimary speakers
18	2.440	0.060	-0.118	Nonprimary speakers
19	3.726	0.041	-0.118	Nonprimary speakers
20	1.837	0.125	-0.284	Nonprimary speakers
21	1.836	0.107	-0.352	Nonprimary speakers
22	1.713	0.102	-0.309	Nonprimary speakers
23	3.884	0.007	-0.146	Nonprimary speakers
24	3.404	0.010	-0.106	n.a.
25	3.463	0.010	-0.140	n.a.

*The Intercept and Score terms were statistically significant at $\alpha = 0.01$ for all items.

**The last column indicates if English Fluency was statistically significant at $\alpha = 0.01$, and if so which group was faster. If the difference between the two groups was not significant, n.a. (not applicable) is indicated.

Table 2. Pearson χ^2 results comparing items not reached with items reached for primary and nonprimary Speakers. Examinees were not matched on test score.

Item Number	Pearson χ^2	Proportion Not Reached	
		Nonprimary Speakers	Primary Speakers
1	.	0	0
2	.	0	0
3	.	0	0
4	.	0	0
5	.	0	0
6	.	.01	0
7	.	.01	0
8	.	.01	0
9	.	.01	0
10	.	.01	0
11	.	.01	0
12	.	.01	0
13	.	.01	0
14	.	.01	0
15	.	.01	0
16	.	.02	.01
17	10.37*	.03	.01
18	11.54*	.03	.01
19	8.78*	.03	.02
20	5.99	.04	.02
21	3.20	.05	.03
22	1.63	.06	.05
23	1.83	.07	.06
24	1.18	.11	.09
25	1.50	.17	.15

*Statistically significant at $\alpha = 0.01$.

Table 3. Pearson χ^2 results comparing rapid guesses and solution behavior for primary and nonprimary Speakers. Examinees were matched on test score.

Item Number	Pearson χ^2	Proportion of Rapid Guesses	
		Nonprimary Speakers	Primary Speakers
1	.	0	0
2	.	0	0
3	.89	.06	.07
4	.	0	0
5	.63	.03	.02
6	.04	.03	.03
7	.	0	0
8	.	0	0
9	17.83*	.05	.01
10	.27	.04	.03
11	0	.02	.02
12	.25	.08	.07
13	.24	.08	.08
14	.04	.12	.12
15	.92	.12	.10
16	4.14	.16	.11
17	3.58	.23	.18
18	6.69*	.29	.21
19	10.85*	.26	.17
20	18.32*	.44	.31
21	17.53*	.50	.36
22	7.68*	.58	.49
23	24.95*	.27	.14
24	7.28*	.35	.27
25	6.03	.41	.33

*Statistically significant at $\alpha = 0.01$.

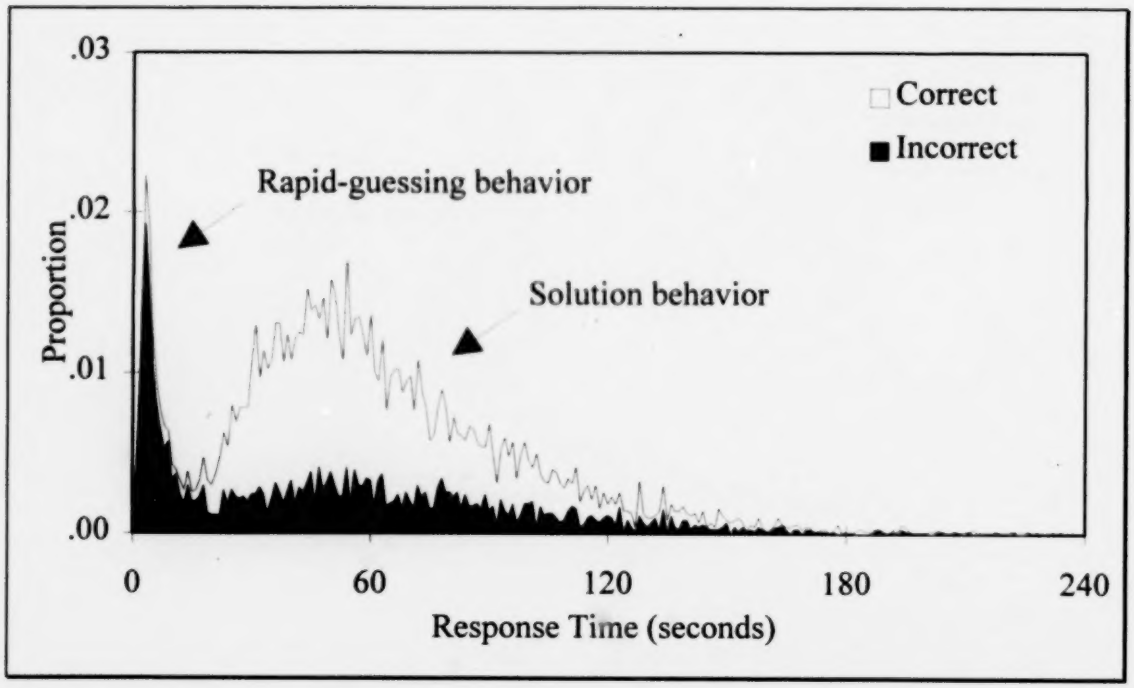

Figure 1. Sample item response time density function. Responses are separated by accuracy for illustrative purposes. (Correct responses are stacked on top of incorrect responses, so the shape of the total distribution is indicated by the top curve.)

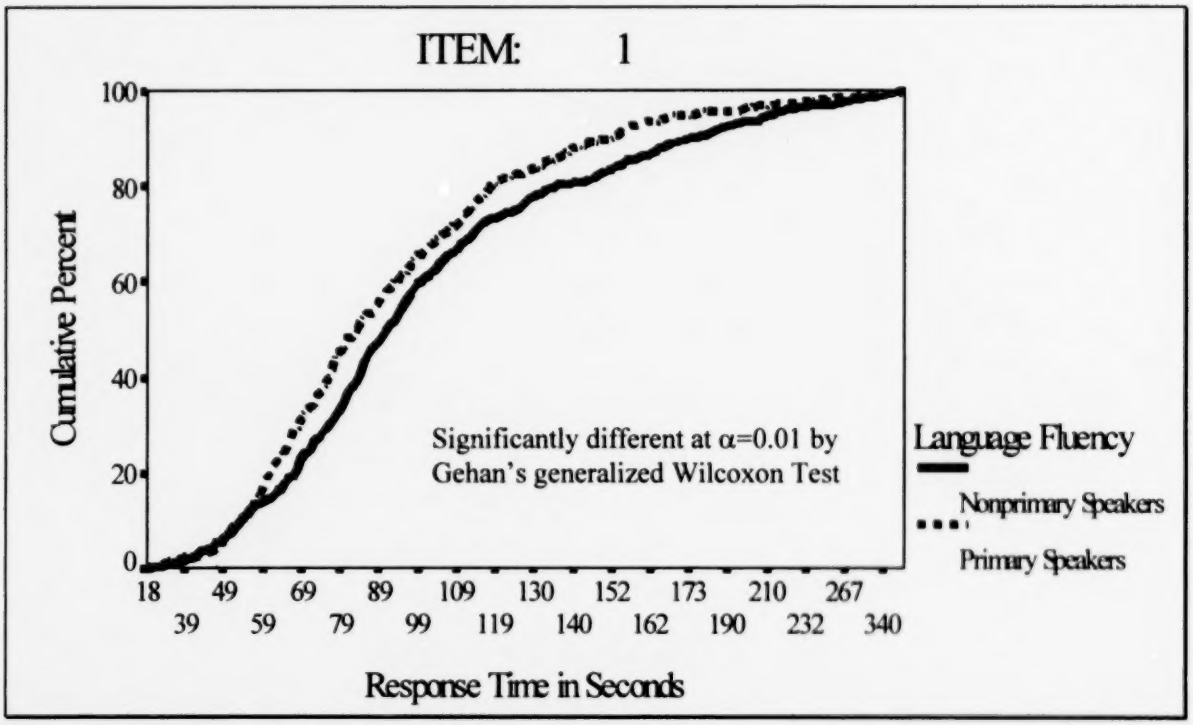

Figure 2. Cumulative distribution function of response times on item 1 for primary and nonprimary speakers of English. The differences between the distributions is significant at $\alpha = 0.01$ by Gehan's generalized Wilcoxon test.

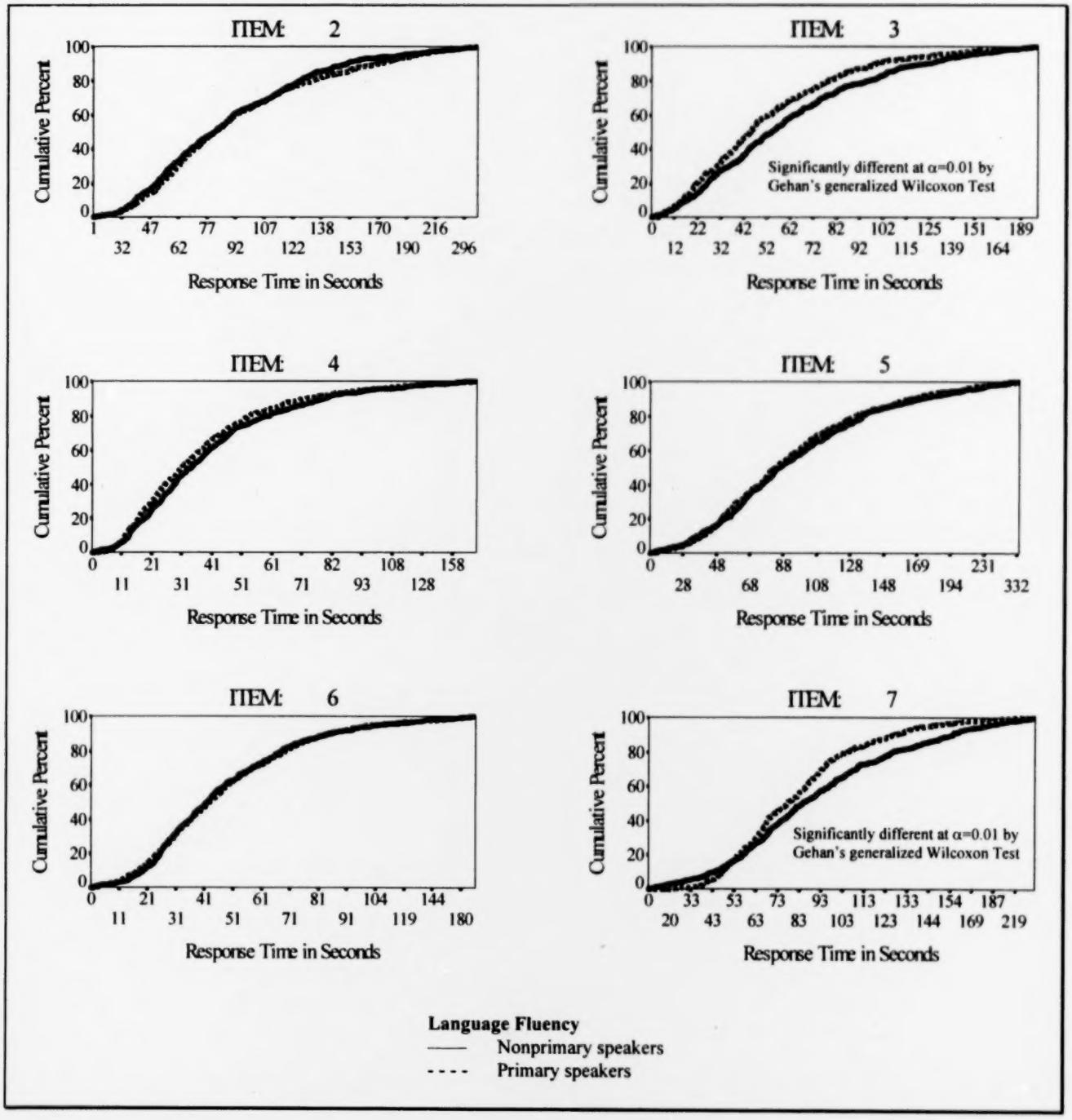
ITEM: 2
Cumulative Percent
100
80
60
40
20
0
1 47 77 107 138 170 216
32 62 92 122 153 190 296
Response Time in Seconds
ITEM: 3
Cumulative Percent
Significantly different at α=0.01 by Gehan's generalized Wilcoxon Test
0 22 42 62 82 102 125 151 189
12 32 52 72 92 115 139 164
Response Time in Seconds
ITEM: 4
Cumulative Percent
0 21 41 61 82 108 158
11 31 51 71 93 128
Response Time in Seconds
ITEM: 5
Cumulative Percent
0 48 88 128 169 231
28 68 108 148 194 332
Response Time in Seconds
ITEM: 6
Cumulative Percent
0 21 41 61 81 104 144
11 31 51 71 91 119 180
Response Time in Seconds
ITEM: 7
Cumulative Percent
Significantly different at α=0.01 by Gehan's generalized Wilcoxon Test
0 33 53 73 93 113 133 154 187
20 43 63 83 103 123 144 169 219
Response Time in Seconds
Language Fluency
Nonprimary speakers
Primary speakers

Figure 3. Cumulative distributions functions (CDFs) of response times for each item for primary and nonprimary speakers of English. CDFs that are significantly different at $\alpha = 0.01$ by Gehan's generalized Wilcoxon are marked as such.

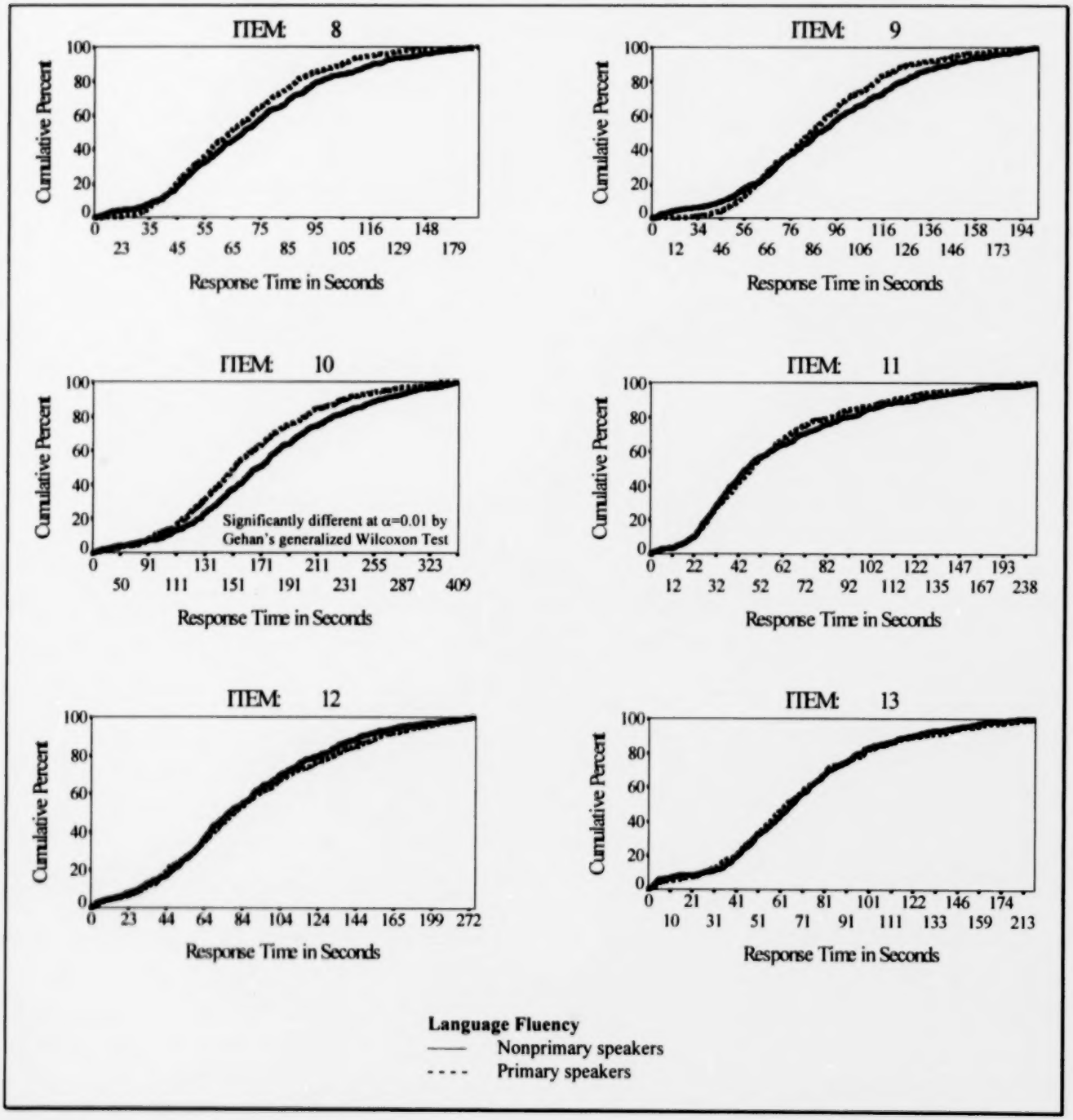

Figure 3, continued. Cumulative distributions functions (CDFs) of response times for each item for primary and nonprimary speakers of English. CDFs that are significantly different at $\alpha = 0.01$ by Gehan's generalized Wilcoxon are marked as such.

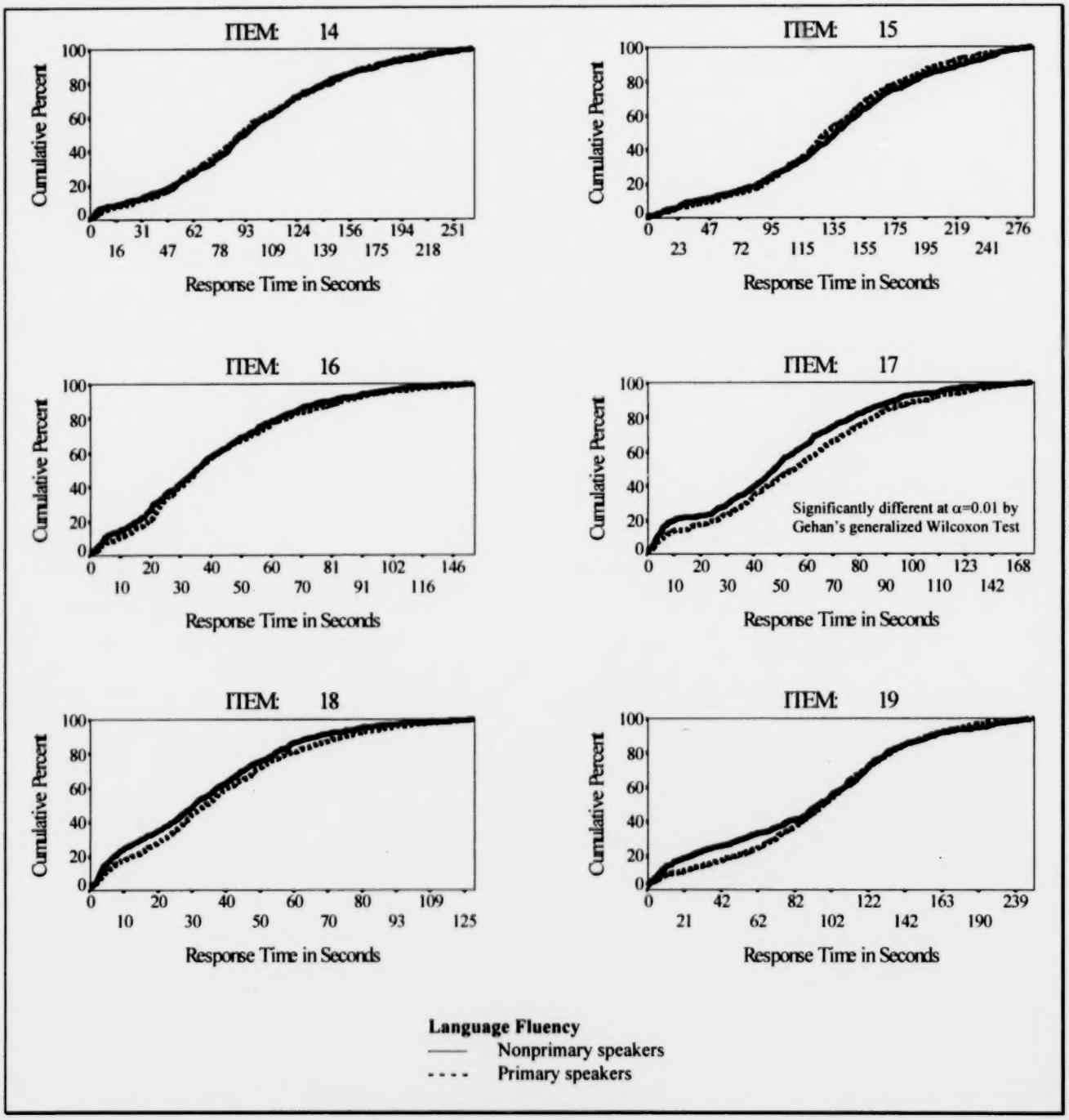

Figure 3, continued. Cumulative distributions functions (CDFs) of response times for each item for primary and nonprimary speakers of English. CDFs that are significantly different at $\alpha = 0.01$ by Gehan's generalized Wilcoxon are marked as such.

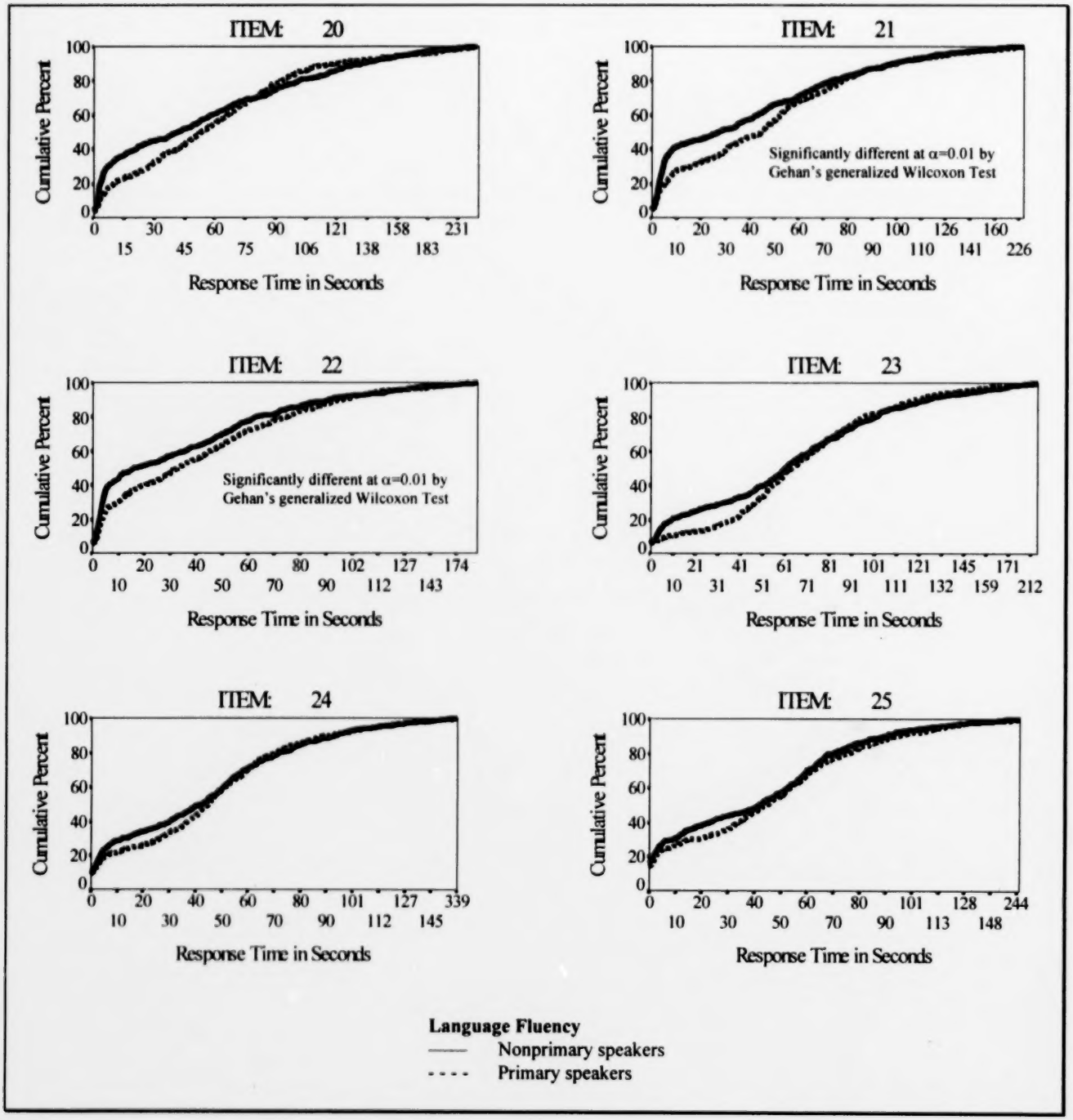
ITEM: 20
Cumulative Percent
100
80
60
40
20
0
0 30 60 90 121 158 231
15 45 75 106 138 183
Response Time in Seconds
ITEM: 21
Cumulative Percent
Significantly different at α=0.01 by Gehan's generalized Wilcoxon Test
0 20 40 60 80 100 126 160
10 30 50 70 90 110 141 226
Response Time in Seconds
ITEM: 22
Cumulative Percent
Significantly different at α=0.01 by Gehan's generalized Wilcoxon Test
0 20 40 60 80 102 127 174
10 30 50 70 90 112 143
Response Time in Seconds
ITEM: 23
Cumulative Percent
0 21 41 61 81 101 121 145 171
10 31 51 71 91 111 132 159 212
Response Time in Seconds
ITEM: 24
Cumulative Percent
0 20 40 60 80 101 127 339
10 30 50 70 90 112 145
Response Time in Seconds
ITEM: 25
Cumulative Percent
0 20 40 60 80 101 128 244
10 30 50 70 90 113 148
Response Time in Seconds
Language Fluency
Nonprimary speakers
Primary speakers

Figure 3, continued. Cumulative distributions functions (CDFs) of response times for each item for primary and nonprimary speakers of English. CDFs that are significantly different at $\alpha = 0.01$ by Gehan's generalized Wilcoxon are marked as such.